



Heriot-Watt University
Research Gateway

Multi-level Monte Carlo methods for the approximation of invariant measures of stochastic differential equations

Citation for published version:

Giles, MB, Majka, MB, Szpruch, L, Vollmer, SJ & Zygalkis, KC 2020, 'Multi-level Monte Carlo methods for the approximation of invariant measures of stochastic differential equations', *Statistics and Computing*, vol. 30, no. 3, pp. 507-524. <https://doi.org/10.1007/s11222-019-09890-0>

Digital Object Identifier (DOI):

[10.1007/s11222-019-09890-0](https://doi.org/10.1007/s11222-019-09890-0)

Link:

[Link to publication record in Heriot-Watt Research Portal](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Statistics and Computing

Publisher Rights Statement:

© The Author(s) 2019

General rights

Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Multi-level Monte Carlo methods for the approximation of invariant measures of stochastic differential equations

Michael B. Giles¹ · Mateusz B. Majka² · Lukasz Szpruch³ · Sebastian J. Vollmer² · Konstantinos C. Zygalakis³

Received: 18 February 2019 / Accepted: 13 August 2019 / Published online: 10 September 2019
 © The Author(s) 2019

Abstract

We develop a framework that allows the use of the multi-level Monte Carlo (MLMC) methodology (Giles in *Acta Numer.* 24:259–328, 2015. <https://doi.org/10.1017/S096249291500001X>) to calculate expectations with respect to the invariant measure of an ergodic SDE. In that context, we study the (over-damped) Langevin equations with a strongly concave potential. We show that when appropriate contracting couplings for the numerical integrators are available, one can obtain a uniform-in-time estimate of the MLMC variance in contrast to the majority of the results in the MLMC literature. As a consequence, a root mean square error of $\mathcal{O}(\varepsilon)$ is achieved with $\mathcal{O}(\varepsilon^{-2})$ complexity on par with Markov Chain Monte Carlo (MCMC) methods, which, however, can be computationally intensive when applied to large datasets. Finally, we present a multi-level version of the recently introduced stochastic gradient Langevin dynamics method (Welling and Teh, in: *Proceedings of the 28th ICML*, 2011) built for large datasets applications. We show that this is the first stochastic gradient MCMC method with complexity $\mathcal{O}(\varepsilon^{-2} |\log \varepsilon|^3)$, in contrast to the complexity $\mathcal{O}(\varepsilon^{-3})$ of currently available methods. Numerical experiments confirm our theoretical findings.

Keywords Numerical analysis · Monte Carlo methods · Stochastic Gradient methods

Mathematics Subject Classification 62F15 · 65N75 · 65L20

1 Introduction

We consider a probability measure $\pi(dx)$ with a density $\pi(x) \propto e^{U(x)}$ on \mathbb{R}^d with an unknown normalising constant. A typical task is the computation of the following quantity

✉ Konstantinos C. Zygalakis
 K.Zygalakis@ed.ac.uk

Michael B. Giles
 Mike.Giles@maths.ox.ac.uk

Mateusz B. Majka
 Mateusz.Majka@warwick.ac.uk

Lukasz Szpruch
 L.Szpruch@ed.ac.uk

Sebastian J. Vollmer
 svollmer@turing.ac.uk

¹ Mathematical Institute, University of Oxford, Oxford, UK

² Department of Statistics, University of Warwick, Coventry, UK

³ School of Mathematics, University of Edinburgh, Edinburgh, UK

$$\pi(g) := \mathbb{E}_\pi g = \int_{\mathbb{R}^d} g(x) \pi(dx), \quad g \in L^1(\pi). \quad (1)$$

Even if $\pi(dx)$ is given in an explicit form, quadrature methods, in general, are inefficient in high dimensions. On the other hand, probabilistic methods scale very well with the dimension and are often the method of choice. With this in mind, we explore the connection between dynamics of stochastic differential equations (SDEs)

$$dX_t = \nabla U(X_t) dt + \sqrt{2} dW_t, \quad X_0 \in \mathbb{R}^d, \quad (2)$$

and the target probability measure $\pi(dx)$. The key idea is that under appropriate assumptions on $U(x)$ one can show that the solution to (2) is ergodic and has $\pi(dx)$ as its unique invariant measure (Has'minskiĭ 1980). However, there exist only a limited number of cases where analytical solutions to (2) are available and typically some form of approximation needs to be employed.

The numerical analysis approach (Kloeden and Platen 1992) is to discretise (2) and run the corresponding Markov chain for a long time interval. One drawback of the numerical

analysis approach is that it might be the case that even though (2) is geometrically ergodic, the corresponding numerical discretisation might not be (Roberts and Tweedie 1996), while in addition extra care is required when ∇U is not globally Lipschitz (Mattingly et al. 2002; Talay 2002; Roberts and Tweedie 1996; Shardlow and Stuart 2000; Hutzenthaler et al. 2011). The numerical analysis approach also introduces bias because the numerical invariant measure does not coincide with the exact one in general (Abdulle et al. 2014; Talay and Tubaro 1990), resulting hence in a biased estimation of $\pi(g)$ in (1). Furthermore, if one uses the Euler–Maruyama method to discretise (2), then computational complexity¹ of $\mathcal{O}(\varepsilon^{-3})$ is required for achieving a root mean square error of order $\mathcal{O}(\varepsilon)$ in the approximation of (1). Furthermore, even if one mitigates the bias due to numerical discretisation by a series of decreasing time steps in combination with an appropriate weighted time average of the quantity of interest (Lamberton and Pagès 2002), the computational complexity still remains $\mathcal{O}(\varepsilon^{-3})$ (Teh et al. 2016).

An alternative way of sampling from $\pi(dx)$ exactly, so that it does not face the bias issue introduced by pure discretisation of (2), is by using the Metropolis–Hastings algorithm (Hastings 1970). We will refer to this as the computational statistics approach. The fact that the Metropolis Hastings algorithm leads to asymptotically unbiased samples of the probability measure is one of the reasons why it has been the method of choice in computational statistics. Moreover, unlike the numerical analysis approach, computational complexity of $\mathcal{O}(\varepsilon^{-2})$ is now required for achieving root mean square error of order $\mathcal{O}(\varepsilon)$ in the (asymptotically unbiased) approximation of (1). We notice that MLMC (Giles 2015) and the unbiaseding scheme (Rhee and Glynn 2012, 2015; Glynn and Rhee 2014) are able to achieve the $\mathcal{O}(\varepsilon^{-2})$ complexity for computing expectations of SDEs on a fixed time interval $[0, T]$, despite using biased numerical discretisations. We are interested in extending this approach to the case of ergodic SDEs on the time interval $[0, \infty)$; see also discussion in Giles (2015).

A particular application of (2) is when one is interested in approximating the posterior expectations for a Bayesian inference problem. More precisely, if for a fixed parameter x the data $\{y_i\}_{i=1,\dots,N}$ are i.i.d. with densities $\pi(y_i|x)$, then $\nabla U(x)$ becomes

$$\nabla U(x) = \nabla \log \pi_0(x) + \sum_{i=1}^N \nabla \log \pi(y_i|x), \quad (3)$$

with $\pi_0(x)$ being the prior distribution of x . When dealing with problems where the number of data items $N \gg 1$ is

¹ In this paper the computational complexity is measured in terms of the expected number of random number generations and arithmetic operations.

large, both the standard numerical analysis and the MCMC approaches suffer due to the high computational cost associated with calculating the likelihood terms $\nabla \log \pi(y_i|x)$ over each data item y_i . One way to circumvent this problem is the stochastic gradient Langevin dynamics algorithm (SGLD) introduced in Welling and Teh (2011), which replaces the sum of the N likelihood terms by an appropriately reweighted sum of $s \ll N$ terms. This leads to the following recursion formula

$$x_{k+1} = x_k + h \left(\nabla \log \pi_0(x_k) + \frac{N}{s} \sum_{i=1}^s \nabla \log \pi(y_{\tau_i^k}|x_k) \right) + \sqrt{2h} \xi_k \quad (4)$$

where ξ_k is a standard Gaussian random variable on \mathbb{R}^d and $\tau^k = (\tau_1^k, \dots, \tau_s^k)$ is a random subset of $[N] = \{1, \dots, N\}$, generated, for example, by sampling with or without replacement from $[N]$. Notice that this corresponds to a noisy Euler discretisation, which for fixed N, s still has computational complexity $\mathcal{O}(\varepsilon^{-3})$ as discussed in Teh et al. (2016) and Vollmer et al. (2016). In this article, we are able to show that careful coupling between fine and coarse paths allows the application of the MLMC framework and hence reduction of the computational complexity of the algorithm to $\mathcal{O}(\varepsilon^{-2} |\log \varepsilon|^3)$. We also remark that coupling in time has been recently further developed in Fang and Giles (2016), Fang and Giles (2017) and Fang and Giles (2019) for Euler schemes.

We would like to stress that in our analysis of the computational complexity of MLMC for SGLD, we treat N and s as fixed parameters. Hence, our results show that in cases in which one is forced to consider $s \ll N$ samples (e.g. in the big data regime, where the cost of taking into account all N samples is prohibitively large) MLMC for SGLD can indeed reduce the computational complexity in comparison with the standard MCMC. However, recently the authors of Nagapetyan et al. (2017) have argued that for the standard MCMC the gain in complexity of SGLD due to the decreased number of samples can be outweighed by the increase in the variance caused by subsampling. We believe that an analogous analysis for MLMC would be highly non-trivial and we leave it for future work.

In summary, the main contributions of this paper are:

1. Extension of the MLMC framework to the time interval $[0, \infty)$ for (2) when U is strongly concave.
2. A convergence theorem that allows the estimation of the MLMC variance using uniform-in-time estimates in the 2-Wasserstein metric for a variety of different numerical methods.
3. A new method of estimation of expectations with respect to the invariant measures without the need of accept/reject

steps (as in MCMC). The methods we propose can be better parallelised than MCMC, since computations on all levels can be performed independently.

4. The application of this scheme to stochastic gradient Langevin dynamics (SGLD) which reduces the complexity of $\mathcal{O}(\varepsilon^{-3})$ to $\mathcal{O}(\varepsilon^{-2} |\log \varepsilon|^3)$ much closer to the standard $\mathcal{O}(\varepsilon^{-2})$ complexity of MCMC.

The rest of the paper is organised as follows: In Sect. 2 we describe the standard MLMC framework, discuss the contracting properties of the true trajectories of (2) and describe an algorithm for applying MLMC with respect to time T for the true solution of (2). In Sect. 3 we present the new algorithm, as well as a framework that allows proving its convergence properties for a numerical method of choice. In Sect. 4 we present two examples of suitable numerical methods, while in Sect. 5 we describe a new version of SGLD with complexity $\mathcal{O}(\varepsilon^{-2} |\log \varepsilon|^3)$. We conclude in Sect. 6 where a number of relevant numerical experiments are described.

2 Preliminaries

In Sect. 2.1 we review the classic, finite time, MLMC framework, while in Sect. 2.2 we state the key asymptotic properties of solutions of (2) when U is strongly concave.

2.1 MLMC with fixed terminal time

Fix $T > 0$ and consider the problem of approximating $\mathbb{E}[g(X_T)]$ where X_T is a solution of the SDE (2) and $g : \mathbb{R}^d \rightarrow \mathbb{R}$. A classical approach to this problem consists of constructing a biased (bias arising due to time-discretisation) estimator of the form

$$\frac{1}{N} \sum_{i=1}^N g((x_T^M)^{(i)}), \quad (5)$$

where $(x_T^M)^{(i)}$ for $i = 1, \dots, N$ are independent copies of the random variable x_T^M , with $\{x_{kh}^M\}_{k=0}^M$ being a discrete time approximation of (2) over $[0, T]$ with the discretisation parameter h and with M time steps, i.e. $Mh = T$. A central limit theorem for the estimator (5) has been derived in Duffie and Glynn (1995), and it was shown that its computational complexity is $\mathcal{O}(\varepsilon^{-3})$, for the root mean square error $\mathcal{O}(\varepsilon)$ (as opposed to $\mathcal{O}(\varepsilon^{-2})$ that can be obtained if we could sample X_T without the bias). The recently developed MLMC approach allows recovering optimal complexity $\mathcal{O}(\varepsilon^{-2})$, despite the fact that the estimator used therein builds on biased samples. This is achieved by exploiting the following identity (Giles 2015; Kebaier 2005)

$$\mathbb{E}[g_L] = \mathbb{E}[g_0] + \sum_{\ell=1}^L \mathbb{E}[g_\ell - g_{\ell-1}], \quad (6)$$

where $g_\ell := g(x_T^{M_\ell})$ and for any $\ell = 0 \dots L$ the Markov chain $\{x_{kh_\ell}^{M_\ell}\}_{k=0}^{M_\ell}$ is the discrete time approximation of (2) over $[0, T]$, with the discretisation parameter h_ℓ and with M_ℓ time steps (hence $M_\ell h_\ell = T$). This identity leads to an unbiased estimator of $\mathbb{E}[g_L]$ given by

$$\frac{1}{N_0} \sum_{i=1}^{N_0} g_0^{(i,0)} + \sum_{\ell=1}^L \left\{ \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} (g_\ell^{(i,\ell)} - g_{\ell-1}^{(i,\ell)}) \right\},$$

where $g_\ell^{(i,\ell)} = g((x_T^{M_\ell})^{(i)})$ and $g_{\ell-1}^{(i,\ell)} = g((x_T^{M_{\ell-1}})^{(i)})$ are independent samples at level ℓ . The inclusion of the level ℓ in the superscript (i, ℓ) indicates that independent samples are used at each level ℓ . The efficiency of MLMC lies in the coupling of $g_\ell^{(i,\ell)}$ and $g_{\ell-1}^{(i,\ell)}$ that results in small $\text{Var}[g_\ell - g_{\ell-1}]$. In particular, for the SDE (2) one can use the same Brownian path to simulate g_ℓ and $g_{\ell-1}$ which, through the strong convergence property of the underlying numerical scheme used, yields an estimate for $\text{Var}[g_\ell - g_{\ell-1}]$.

By solving a constrained optimisation problem (cost and accuracy), one can see that reduced computational complexity (variance) arises since the MLMC method allows one to efficiently combine many simulations on low accuracy grids (at a corresponding low cost), with relatively few simulations computed with high accuracy and high cost on very fine grids. It is shown in Giles (2015) that under the assumptions²

$$|\mathbb{E}[g - g_\ell]| \leq c_1 h_\ell^\alpha, \quad \text{Var}[g_\ell - g_{\ell-1}] \leq c_2 h_\ell^\beta, \quad (7)$$

for some $\alpha \geq 1/2$, $\beta > 0$, $c_1 > 0$ and $c_2 > 0$, the computational complexity of the resulting multi-level estimator with accuracy ε is proportional to

$$C = \begin{cases} \varepsilon^{-2}, & \beta > \gamma, \\ \varepsilon^{-2} \log^2(\varepsilon), & \beta = \gamma, \\ \varepsilon^{-2-(1-\beta)/\alpha}, & 0 < \beta < \gamma \end{cases}$$

where the cost of the algorithm is of order $h_L^{-\gamma}$. Typically, the constants c_1, c_2 grow exponentially in time T as they follow from classical finite time weak and strong convergence analysis of the numerical schemes. The aim of this paper is to establish the bounds (7) uniformly in time, i.e. to find constants $\tilde{c}_1, \tilde{c}_2 > 0$ independent of T such that

$$\sup_{T>0} |\mathbb{E}[g - g_\ell]| \leq \tilde{c}_1 h_\ell^\alpha, \quad \sup_{T>0} \text{Var}[g_\ell - g_{\ell-1}] \leq \tilde{c}_2 h_\ell^\beta. \quad (8)$$

² Recall h_ℓ is the time step used in the discretisation of the level ℓ .

Remark 2.1 The reader may notice that in the regime when $\beta > \gamma$, the computational complexity of $\mathcal{O}(\varepsilon^{-2})$ coincides with that of an unbiased estimator. Nevertheless, the MLMC estimator as defined here is still biased, with the bias being controlled by the choice of final level parameter L . However, in this setting it is possible to eliminate the bias by a clever randomisation trick (Rhee and Glynn 2015).

2.2 Properties of ergodic SDEs with strongly concave drifts

Consider the SDE (2) and let U satisfy the following condition

HU0 For any $x, y \in \mathbb{R}^d$ there exists a positive constant m s.t.

$$\langle \nabla U(y) - \nabla U(x), y - x \rangle \leq -m|x - y|^2, \quad (9)$$

which is also known as a one-side Lipschitz condition. Condition **HU0** is satisfied for strongly concave potential, i.e. when for any $x, y \in \mathbb{R}^d$ there exists constant m s.t.

$$U(y) \leq U(x) + \langle \nabla U(x), y - x \rangle - \frac{m}{2}|x - y|^2.$$

In addition **HU0** implies that

$$\langle \nabla U(x), x \rangle \leq -\frac{m}{2}|x|^2 + \frac{1}{2m}|\nabla U(0)|^2, \quad \forall x \in \mathbb{R}^d \quad (10)$$

which in turn implies that

$$\langle \nabla U(x), x \rangle \leq -m'|x|^2 + 2b|\nabla U(0)|^2, \quad \forall x \in \mathbb{R}^d \quad (11)$$

for some³ $m' > 0, b \geq 0$. Condition **HU0** ensures the contraction needed to establish uniform-in-time estimates for the solutions of (2). For the transparency of the exposition we introduce the following flow notation for the solution to (2), starting at $X_0 = x$

$$\psi_{s,t,W}(x) := x + \int_s^t \nabla U(X_r) dr + \int_s^t \sqrt{2} dW_r, \quad x \in \mathbb{R}^d. \quad (12)$$

The theorem below demonstrates that solutions to (2) driven by the same Brownian motion, but with different initial conditions, enjoy an exponential contraction property.

Theorem 2.2 Let $(W_t)_{t \geq 0}$ be a standard Brownian motion in \mathbb{R}^d . We fix random variables $Y_0, X_0 \in \mathbb{R}^d$ and define $X_T = \psi_{0,T,W}(X_0)$ and $Y_T = \psi_{0,T,W}(Y_0)$. If **HU0** holds, then

³ If $\nabla U(0) = 0$ then $m' = m, b = 0$. Otherwise $m' < m$ (implication of Young's inequality).

$$\mathbb{E}|X_T - Y_T|^2 \leq \mathbb{E}|X_0 - Y_0|^2 e^{-2mT} \quad (13)$$

Proof The result follows from Itô's formula. Indeed, we have

$$\begin{aligned} \frac{1}{2} e^{2mt} |X_t - Y_t|^2 &= \frac{1}{2} |X_0 - Y_0|^2 + \int_0^t m e^{2ms} |X_s - Y_s|^2 ds \\ &\quad + \int_0^t e^{2ms} \langle \nabla U(X_s) - \nabla U(Y_s), X_s - Y_s \rangle ds. \end{aligned}$$

Assumption **HU0** yields

$$\mathbb{E}|X_T - Y_T|^2 \leq e^{-2mT} \mathbb{E}|X_0 - Y_0|^2,$$

as required. \square

Remark 2.3 The 2-Wasserstein distance between probability measures ν_1 and ν_2 defined on a Polish metric space E , is given by

$$\mathcal{W}_2(\nu_1, \nu_2) = \left(\inf_{\pi \in \Gamma(\nu_1, \nu_2)} \int_{E \times E} |x - y|^2 \pi(dx, dy) \right)^{\frac{1}{2}},$$

with $\Gamma(\nu_1, \nu_2)$ being the set of couplings of ν_1 and ν_2 (all probability measures on $E \times E$ with marginals ν_1 and ν_2). We denote $\mathcal{L}(\psi_{0,t,W}(x)) = P_t(x, \cdot)$. That is, P_t is the transition kernel of the SDE (2). Since the choice of the same driving Brownian motion in Theorem 2.2 is an example of a coupling, Equation (13) implies

$$\mathcal{W}_2(P_t(x, \cdot), P_t(y, \cdot)) \leq |x - y| \exp(-mt) \quad (14)$$

Consequently, P_t has a unique invariant measure, and thus, the process is ergodic (Hairer et al. 2011). In the present paper we are not concerned with determining couplings that are optimal; for practical considerations one should only consider couplings that are feasible to implement [see also discussion in Agapiou et al. (2018) and Giles and Szpruch (2014)].

2.3 Coupling in time T

For the MLMC method with different discretisation parameters on different levels, coupling with the same Brownian motion is not enough to obtain good upper bounds on the variance, as, in general, solutions to SDEs (2) are $1/2$ -Hölder continuous (Krylov 2009), i.e. for any $t > s > 0$ there exists a constant $C > 0$ such that

$$\mathbb{E}|X_t - X_s|^2 \leq C|t - s| \quad (15)$$

and it is well known that this bound is sharp. As we shall see later, this bound will not lead to an efficient MLMC implementation. However, by suitable coupling of the SDE

solutions on time intervals of length T and S , $T > S$, respectively, we will be able to take advantage of the exponential contraction property obtained in Theorem 2.2.

Let $(T_\ell)_{\ell \geq 0}$ be an increasing sequence of positive real numbers. To couple processes with different terminal times T_i and T_j , $i \neq j$, we exploit the time homogeneous Markov property of the flow (12). More precisely, for each $\ell \geq 0$ one would like to construct a pair $(\mathcal{X}^{(f,\ell)}, \mathcal{X}^{(c,\ell)})$ of solutions to (2), which we refer to as fine and coarse paths, such that

$$\begin{aligned}\mathcal{L}(\mathcal{X}^{(f,\ell)}(T_{\ell-1})) &= \mathcal{L}(X_{T_\ell}), \\ \mathcal{L}(\mathcal{X}^{(c,\ell)}(T_{\ell-1})) &= \mathcal{L}(X_{T_{\ell-1}}), \quad \forall \ell \geq 0,\end{aligned}\quad (16)$$

and

$$\mathbb{E}|\mathcal{X}^{(f,\ell)}(T_{\ell-1}) - \mathcal{X}^{(c,\ell)}(T_{\ell-1})|^2 \leq \mathbb{E}|X_{T_\ell} - X_{T_{\ell-1}}|^2. \quad (17)$$

Following Rhee and Glynn (2012, 2015); Agapiou et al. (2018); Giles (2015), we propose a particular coupling denoted by $(X^{(f,\ell)}, X^{(c,\ell)})$, and constructed in the following way (see also Fig. 1a)

- First⁴ obtain a solution to (2) over $[0, T_\ell - T_{\ell-1}]$. We fix an \mathbb{R}^d -valued random variable $X(0)$ and take $X^{(f,\ell)}(0) = \psi_{0, (T_\ell - T_{\ell-1}), \tilde{W}}(X(0))$.
- Next couple fine and coarse paths on the remaining time interval $[0, T_{\ell-1}]$ using the same Brownian motion W , i.e.

$$X^{(f,\ell)}(T_{\ell-1}) = \psi_{0, T_{\ell-1}, W}(X^{(f,\ell)}(0)),$$

and

$$X^{(c,\ell)}(T_{\ell-1}) = \psi_{0, T_{\ell-1}, W}(X(0)).$$

We note here that $\nabla U(\cdot)$ in (2) is time homogenous; hence, the same applies for the corresponding transition kernel $\mathcal{L}(\psi_{0,t,W}(x)) = P_t(x, \cdot)$, which implies that condition (16) holds. Now Theorem 2.2 yields

$$\mathbb{E}|X^{(f,\ell)}(T_{\ell-1}) - X^{(c,\ell)}(T_{\ell-1})|^2 \leq \mathbb{E}|X^{(f,\ell)}(0) - X(0)|^2 e^{-2mT_{\ell-1}}. \quad (18)$$

implying that condition (17) is also satisfied. We now take $\rho > 1$ and define

$$T_\ell := \frac{\log 2}{2m} \rho(\ell + 1) \quad \forall \ell \geq 0. \quad (19)$$

In our case $g_\ell^{(i,\ell)} = g((X^{(f,\ell)}(T_{\ell-1}))^{(i)})$ and $g_{\ell-1}^{(i,\ell)} = g((X^{(c,\ell)}(T_{\ell-1}))^{(i)})$ and we assume that g is globally Lipschitz with Lipschitz constant K . Hence,

⁴ As we can see in Fig. 1b, doing this first is important for the overall difference of the paths.

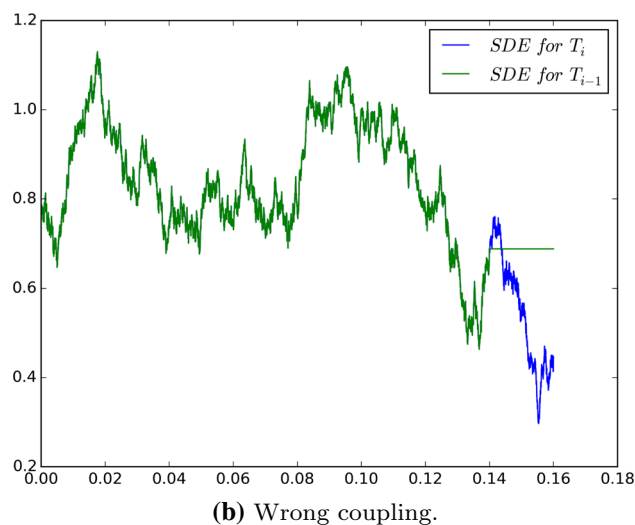
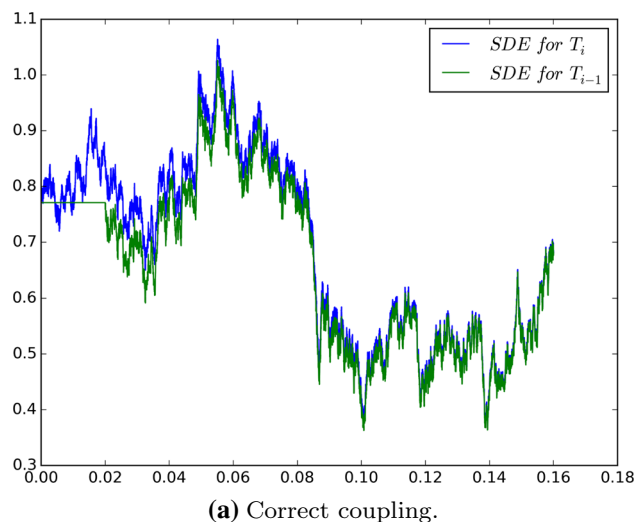


Fig. 1 Shifted couplings

$$\begin{aligned}\mathbb{E}|g(X^{(f,\ell)}(T_{\ell-1})) - g(X^{(c,\ell)}(T_{\ell-1}))|^2 &\leq K^2 \mathbb{E}|X^{(f,\ell)}(T_{\ell-1}) - X^{(c,\ell)}(T_{\ell-1})|^2 \\ &\leq K^2 \mathbb{E}|X^{(f,\ell)}(0) - X(0)|^2 e^{-2mT_{\ell-1}} \\ &\leq K^2 \mathbb{E}|X^{(f,\ell)}(0) - X(0)|^2 2^{-\rho\ell} \\ &\leq K^2 C |T_\ell - T_{\ell-1}| 2^{-\rho\ell}, \quad \forall i \geq 0,\end{aligned}$$

where the last inequality follows from (15).

3 MLMC in T for approximation of SDEs

Having described a coupling algorithm with good contraction properties, we now present the main algorithm in Sect. 3.1, while in Sect. 3.2 we present a general numerical analysis framework for proving the convergence of our algorithm.

3.1 Description of the general algorithm

We now focus on the numerical discretisation of the Langevin equation (2). In particular, we are interested in coupling the discretisations of (2) based on step size h_ℓ and $h_{\ell-1}$ with $h_\ell = h_0 2^{-\ell}$. Furthermore, as we are interested in computing expectations with respect to the invariant measure $\pi(dx)$, we also increase the time endpoint $T_\ell \uparrow \infty$ which is chosen such that $\frac{T_\ell}{h_0}, \frac{T_\ell}{h_\ell} \in \mathbb{N}$. We illustrate the main idea using two generic discrete time stochastic processes $(x_k)_{k \in \mathbb{N}}, (y_k)_{k \in \mathbb{N}}$ which can be defined as

$$x_{k+1} = S_{h,\xi_k}^f(x_k), \quad y_{k+1} = S_{h,\xi_k}^c(y_k), \quad (20)$$

where $S_{h,\xi_k}(x_k) = S(x_k, h, \xi_k)$ and the operators $S^f, S^c : \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}^{d \times m} \rightarrow \mathbb{R}^d$ are Borel measurable, whereas $\xi, \tilde{\xi}$ are random inputs to the algorithms. The operators S^f and S^c in (20) need not be the same. This extra flexibility allows analysing various coupling ideas.

For example, for the Euler discretisation we have

$$S_{h,\xi}(x) = x + h \nabla U(x) + \sqrt{2h} \xi,$$

where $\xi \sim \mathcal{N}(0, I)$. We will also use the notation $P_h(x, \cdot) = \mathcal{L}(S_{h,\xi}(x))$ for the corresponding Markov kernel.

For MLMC algorithms one evolves both fine and coarse paths jointly, over a time interval of length $T_{\ell-1}$, by doing two steps for the finer level (with the time step h_ℓ) and one on the coarser level (with the time step $h_{\ell-1}$). We will use the notation $(x^f), (x^c)$ for

$$x_{k+\frac{1}{2}}^f = S_{\frac{h}{2}, \xi_{k+\frac{1}{2}}}^f(x_k^f), \quad x_{k+1}^f = S_{\frac{h}{2}, \xi_{k+1}}^f(x_{k+\frac{1}{2}}^f) \quad (21)$$

$$x_{k+1}^c = S_{h, \tilde{\xi}_{k+1}}^c(x_k^c). \quad (22)$$

The algorithm generating $(x_k^f)_{k \in \mathbb{N}/2}$ and $(x_k^c)_{k \in \mathbb{N}}$ is presented in Algorithm 1.

3.2 General convergence analysis

We will now present a general theorem for estimating the bias and the variance in the MLMC set up. We refrain from prescribing the exact dynamics of $(x_k)_{k \geq 0}$ and $(y_k)_{k \geq 0}$ in (20), as we seek general conditions allowing the construction of uniform-in-time approximations of (2) in the L^2 -Wasserstein norm. The advantage of working in this general setting is that if we wish to work with more advanced numerical schemes than the Euler method (e.g. implicit, projected, adapted or randomised scheme) or general noise terms (e.g. α -stable processes), it will be sufficient to verify relatively simple conditions to see the performance of the complete algorithm. To

1. Set $x_0^{(f,\ell)} = x_0$, then simulate according to $P_{h_\ell}(x_0, \cdot)$ up to time $\frac{T_\ell - T_{\ell-1}}{h_\ell}$, thus obtaining $x_{\frac{T_\ell - T_{\ell-1}}{h_\ell}}^{(f,\ell)}$;
2. Set $x_0^{(c,\ell)} = x_0$ and $x_0^{(f,\ell)} = x_{\frac{T_\ell - T_{\ell-1}}{h_\ell}}^{(f,\ell)}$, then simulate $(x^{(f,\ell)}, x^{(c,\ell)})$ jointly as

$$\left(x_{k+1}^{(f,\ell)}, x_{k+1}^{(c,\ell)}\right) = \left(S_{h_\ell, \xi_{k+1}}^f \circ S_{h_\ell, \xi_{k+1}}^f(x_k^{(f,\ell)}), S_{h_{\ell-1}, \frac{1}{\sqrt{2}}(\xi_{k+\frac{1}{2}} + \xi_{k+1})}^c(x_k^{(c,\ell)})\right).$$
3. Set $k_\ell := \frac{T_\ell - 1}{h_{\ell-1}}$ and

$$\Delta_\ell^{(i)} := g\left(\left(x_{k_\ell}^{(f,\ell)}\right)^{(i)}\right) - g\left(\left(x_{k_\ell}^{(c,\ell)}\right)^{(i)}\right)$$

Algorithm 1: Coupling Langevin discretisations for $T_\ell \uparrow \infty$.

give the reader some intuition behind the abstract assumptions, we discuss the specific methods in Sect. 4.

3.2.1 Uniform estimates in time

Definition 3.1 (Bias) We say that a process $(x_k)_{k \in \mathbb{N}}$ converges weakly uniformly in time with order $\alpha > 0$ to the solution $(X_t)_{t \geq 0}$ of the SDE (2), if there exists a constant $c > 0$ such that for any $h > 0$,

$$\sup_{t \geq 0} |\mathbb{E}[g(X_t)] - \mathbb{E}[g(x_{\lfloor t/h \rfloor})]| \leq ch^\alpha \quad g \in C_K^r(\mathbb{R}).$$

We define MLMC variance as follows:

Definition 3.2 (MLMC variance) Let the operators in (21)–(22) satisfy that for all x

$$\mathcal{L}\left(S_{h,\xi}^f(x)\right) = \mathcal{L}\left(S_{h,\tilde{\xi}}^c(x)\right). \quad (23)$$

We say that the MLMC variance is of order $\beta > 0$ if there exists a constant $c_V > 0$ s.t. for any $h > 0$,

$$\sup_{t \geq 0} \mathbb{E}|g(x_{\lfloor t/h \rfloor}^c) - g(x_{\lfloor t/h \rfloor}^f)|^2 \leq c_V h^\beta. \quad (24)$$

3.2.2 Statement of sufficient conditions

We now discuss the necessary conditions imposed on a generic numerical method (20) to estimate MLMC variance. We decompose the global error into the one step error and the regularity of the scheme. To proceed we introduce the notation x_{k,x_s}^h for the process at time k with initial condition x_s at time $s < k$. If it is clear from the context what initial condition is used, we just write x_k^h . We also define the

conditional expectation operator as $\mathbb{E}_n[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_n]$, where $\mathcal{F}_n := \sigma(x_k^h : k \leq n)$.

We now have the following definition.

Definition 3.3 (L^2 -regularity). We will say that the one step operator $S : \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}^{d \times m} \rightarrow \mathbb{R}^d$ is L^2 -regular **uniformly in time** if for any \mathcal{F}_n -measurable random variables $x_n, y_n \in \mathbb{R}^d$ there exist constants $K, C_{\mathcal{R}}, C_{\mathcal{H}}, \tilde{\beta} > 0$ and random variables $Z_{n+1}, \mathcal{R}_{n+1} \in \mathcal{F}_{n+1}$ and $\mathcal{H}_n \in \mathcal{F}_n$, such that for all $h \in (0, 1)$

$$S_{h, \xi_{n+1}}(x_n) - S_{h, \xi_{n+1}}(y_n) = x_n - y_n + Z_{n+1}$$

and

$$\begin{aligned} \mathbb{E}_n[|S_{h, \xi_{n+1}}(x_n) - S_{h, \xi_{n+1}}(y_n)|^2] &\leq (1 - Kh)|x_n - y_n|^2 \\ &\quad + \mathcal{R}_n \\ \mathbb{E}_n[|Z_{n+1}|^2] &\leq \mathcal{H}_n|x_n - y_n|^2 h, \end{aligned} \quad (25)$$

where

$$\begin{aligned} \sup_{n \geq 1} \mathbb{E} \left[\sum_{i=0}^{n-1} e^{(i-(n-1))hK/2} \mathcal{R}_i \right] &\leq C_{\mathcal{R}} h^{\tilde{\beta}}, \\ \sup_{n \geq 1} \mathbb{E} [|\mathcal{H}_n|^2] &\leq C_{\mathcal{H}}. \end{aligned} \quad (26)$$

We now introduce the set of the assumptions needed for the proof of the main convergence theorem.

Assumption 1 Consider two processes $(x_k^f)_{2k \in \mathbb{N}}$ and $(x_k^c)_{k \in \mathbb{N}}$ obtained from the recursive application of the operators $S_{h, \xi}^f(\cdot)$ and $S_{h, \xi}^c(\cdot)$ as defined in (20). We assume that

H0 There exists a constant $H > 0$ such that for all $q > 1$

$$\sup_k \mathbb{E}|x_k^f|^q \leq H \quad \text{and} \quad \sup_k \mathbb{E}|x_k^c|^q \leq H.$$

H1 For any $x \in \mathbb{R}^d$

$$\mathcal{L}(S_{h, \xi}^f(x)) = \mathcal{L}(S_{h, \xi}^c(x)).$$

H2 The operator $S_{h, \xi}^f(\cdot)$ is L^2 -regular uniformly in time.

Below, we present the main convergence result of this section. By analogy to (21)–(22), we use the notation

$$\begin{aligned} x_{n, x_{n-1}}^f &= S_{h, \xi_n}^f \left(x_{n-\frac{1}{2}, x_{n-1}}^f \right) \\ x_{n-\frac{1}{2}, x_{n-1}}^f &= S_{h, \xi_{n-\frac{1}{2}}}^f (x_{n-1}^c) \\ x_{n, x_{n-1}}^c &= S_{h, \xi_n}^c (x_{n-1}^c). \end{aligned}$$

Using the estimates derived here, we can immediately estimate the rate of decay of MLMC variance.

Theorem 3.4 Take $(x_n^f)_{n \in \mathbb{N}/2}$ and $(x_n^c)_{n \in \mathbb{N}}$ with $h \in (0, 1]$ and assume that **H0–H2** hold. Moreover, assume that there exist constants $c_s > 0, c_w > 0$ and $\alpha \geq \frac{1}{2}, \beta \geq 0, p \geq 1$ with $\alpha \geq \frac{\beta}{2}$ such that for all $n \geq 1$

$$|\mathbb{E}_{n-1}(x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f)| \leq c_w(1 + |x_{n-1}^c|^p)h^{\alpha+1}, \quad (27)$$

and

$$\mathbb{E}_{n-1}[|x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f|^2] \leq c_s(1 + |x_{n-1}^c|^{2p})h^{\beta+1}. \quad (28)$$

Then, the global error is bounded by

$$\begin{aligned} \mathbb{E}[|x_{T/h, x_0}^c - x_{T/h, y_0}^f|^2] &\leq |x_0 - y_0|^2 e^{-K/2T} + 2\Gamma h^\beta / K \\ &\quad + \sum_{j=0}^{n-1} e^{(j-(n-1))hK/2} \mathbb{E}(\mathcal{R}_j), \end{aligned}$$

where $T/h = n$ and Γ is given by (29).

Proof We begin using the following identity

$$\begin{aligned} x_{n, x_0}^c - x_{n, y_0}^f &= x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f \\ &= (x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f) + (x_{n, x_{n-1}}^f - x_{n, x_{n-1}}^f). \end{aligned}$$

We will be able to deal with the first term in the sum by using Equations (27) and (28), while the second term will be controlled because of the L^2 regularity of the numerical scheme. Indeed, by squaring both sides in the equality above, we have

$$\begin{aligned} |x_{n, y_0}^c - x_{n, x_0}^f|^2 &= |x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f|^2 \\ &\quad + |x_{n, x_{n-1}}^f - x_{n, x_{n-1}}^f|^2 \\ &\quad + 2(x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f, x_{n-1}^c - x_{n-1}^f + Z_n), \end{aligned}$$

where in the last line we have used Assumption **H2**. Applying conditional expectation operator to both sides of the above equality, we obtain

$$\begin{aligned} \mathbb{E}_{n-1}[|x_{n, y_0}^c - x_{n, x_0}^f|^2] &= \mathbb{E}_{n-1}[|x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f|^2] \\ &\quad + \mathbb{E}_{n-1}[|x_{n, x_{n-1}}^f - x_{n, x_{n-1}}^f|^2] \\ &\quad + 2(x_{n-1}^c - x_{n-1}^f, \mathbb{E}_{n-1}[x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f]) \\ &\quad + 2\mathbb{E}_{n-1}\langle Z_n, x_{n, x_{n-1}}^c - x_{n, x_{n-1}}^f \rangle. \end{aligned}$$

Applying Cauchy–Schwarz inequality and using the weak error estimate (27) lead to

$$\begin{aligned} \mathbb{E}_{n-1}[|x_{n,y_0}^c - x_{n,x_0}^f|^2] &\leq \mathbb{E}_{n-1}[|x_{n,x_{n-1}}^c - x_{n,x_{n-1}}^f|^2] \\ &\quad + \mathbb{E}_{n-1}[|x_{n,x_{n-1}}^f - x_{n,x_{n-1}}^f|^2] \\ &\quad + 2c_w h^{\alpha+1} |x_{n-1}^c - x_{n-1}^f| (1 + |x_{n-1}^c|^p) \\ &\quad + 2(\mathbb{E}_{n-1}[|Z_n|^2])^{1/2} (\mathbb{E}_{n-1}[|x_{n,x_{n-1}}^c - x_{n,x_{n-1}}^f|^2])^{1/2}. \end{aligned}$$

By Assumptions **H0–H2** and the strong error estimate (28), we have

$$\begin{aligned} \mathbb{E}_{n-1}[|x_{n,y_0}^c - x_{n,x_0}^f|^2] &\leq c_s(1 + |x_{n-1}^c|^{2p})h^{\beta+1} \\ &\quad + |x_{n-1}^c - x_{n-1}^f|^2(1 - Kh) + \mathcal{R}_{n-1} \\ &\quad + 2c_w h^{\alpha+1} |x_{n-1}^c - x_{n-1}^f| (1 + |x_{n-1}^c|^p) \\ &\quad + 2(\mathbb{E}_{n-1}[\mathcal{H}_n] |x_{n-1}^c - x_{n-1}^f|^2 h)^{1/2} (c_s(1 + |x_{n-1}^c|^{2p})h^{\beta+1})^{1/2} \\ &\leq c_s(1 + |x_{n-1}^c|^{2p})h^{\beta+1} + |x_{n-1}^c - x_{n-1}^f|^2(1 - Kh) + \mathcal{R}_{n-1} \\ &\quad + 2c_w h^{\alpha+1} |x_{n-1}^c - x_{n-1}^f| (1 + |x_{n-1}^c|^p) \\ &\quad + 2(|x_{n-1}^c - x_{n-1}^f|^2 h)^{1/2} (c_s \mathbb{E}_{n-1}[\mathcal{H}_n(1 + |x_{n-1}^c|^{2p})]h^{\beta+1})^{1/2}, \end{aligned}$$

while taking expected values and applying Cauchy–Schwarz inequality and the fact that $\alpha \geq \frac{\beta}{2}$ and $h < 1$ (and hence $h^{\alpha+1} \leq h^{\frac{\beta}{2}+1}$) give

$$\begin{aligned} \mathbb{E}[|x_{n,y_0}^c - x_{n,x_0}^f|^2] &\leq c_s(1 + \mathbb{E}[|x_{n-1}^c|^{2p}])h^{\beta+1} \\ &\quad + \mathbb{E}[|x_{n-1}^c - x_{n-1}^f|^2](1 - Kh) + \mathbb{E}[\mathcal{R}_{n-1}] \\ &\quad + 2\sqrt{2}c_w (\mathbb{E}[|x_{n-1}^c - x_{n-1}^f|^2 h])^{1/2} (\mathbb{E}[(1 + |x_{n-1}^c|^{2p})h^{\beta+1}])^{1/2} \\ &\quad + 2\mathbb{E}[|x_{n-1}^c - x_{n-1}^f|^2 h]^{1/2} \mathbb{E}[c_s \mathcal{H}_{n-1}(1 + |x_{n-1}^c|^{2p})h^{\beta+1}]^{1/2}. \end{aligned}$$

Now Young's inequality gives that for any $\varepsilon > 0$

$$\begin{aligned} \mathbb{E}[|x_{n-1}^c - x_{n-1}^f|^2 h]^{1/2} \mathbb{E}[(1 + |x_{n-1}^c|^{2p})h^{\beta+1}]^{1/2} \\ \leq \varepsilon \mathbb{E}[(x_{n-1}^c - x_{n-1}^f)^2] h + \frac{1}{4\varepsilon} \mathbb{E}[(1 + |x_{n-1}^c|^{2p})] h^{\beta+1} \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[|x_{n-1}^c - x_{n-1}^f|^2 h]^{1/2} \mathbb{E}[c_s \mathcal{H}_{n-1}(1 + |x_{n-1}^c|^{2p})h^{\beta+1}]^{1/2} \\ \leq \varepsilon \mathbb{E}[|x_{n-1}^c - x_{n-1}^f|^2] h + \frac{1}{4\varepsilon} \mathbb{E}[c_s \mathcal{H}_{n-1}(1 + |x_{n-1}^c|^{2p})] h^{\beta+1}, \end{aligned}$$

while

$$\mathbb{E}[\mathcal{H}_{n-1}(1 + |x_{n-1}^c|^{2p})] \leq \frac{1}{2} \mathbb{E}[\mathcal{H}_{n-1}^2] + \mathbb{E}[(1 + |x_{n-1}^c|^{4p})].$$

Let $\gamma_n := \mathbb{E}[|x_{n,y_0}^c - x_{n,x_0}^f|^2]$. Since $(1 + \mathbb{E}[|x_{n-1}^c|^{2p}]) \leq (1 + \mathbb{E}[|x_{n-1}^c|^{4p}])$, we have

$$\begin{aligned} \gamma_n &\leq \left(c_s H + \frac{2\sqrt{2}c_w H + c_s(\mathbb{E}[\mathcal{H}_{n-1}^2] + 2H)}{4\varepsilon} \right) h^{\beta+1} \\ &\quad + \mathbb{E}[\mathcal{R}_{n-1}] + \gamma_{n-1}(1 - [K - (2\sqrt{2}c_w + 2)\varepsilon]h) \end{aligned}$$

Fix $\varepsilon = \frac{K}{2(2\sqrt{2}c_w + 2)}$, and define

$$\begin{aligned} \Gamma &:= \left(c_s H + (2\sqrt{2}c_w + 2) \right. \\ &\quad \times \left. \frac{(2\sqrt{2}c_w H + c_s(\mathbb{E}[\sup_n |\mathcal{H}_{n-1}|^2] + 2H))}{2K} \right). \end{aligned} \quad (29)$$

We have

$$\gamma_n \leq (1 - Kh/2) \gamma_{n-1} + \Gamma h^{\beta+1} + \mathbb{E}[\mathcal{R}_{n-1}]. \quad (30)$$

We complete the proof by Lemma 3.5. \square

Lemma 3.5 Let $a_n, g_n, c \geq 0$, $n \in \mathbb{N}$ be given. Moreover, assume that $1 + \lambda > 0$. Then, if $a_n \in \mathbb{R}$, $n \in \mathbb{N}$, satisfies

$$a_{n+1} \leq a_n(1 + \lambda) + g_n + c, \quad n \geq 0,$$

then

$$a_n \leq a_0 e^{n\lambda} + c \frac{e^{n\lambda} - 1}{\lambda} + \sum_{j=0}^{n-1} g_j e^{((n-1)-j)\lambda}, \quad n \geq 1.$$

Remark 3.6 Note that if we can choose $\tilde{\beta} > \beta$ in (26) (which, as we will see in Sect. 4, is the case, for example, for Euler and implicit Euler schemes), then from Theorem 3.4 we get

$$\mathbb{E}[|x_{T/h,x_0}^c - x_{T/h,y_0}^f|^2] \leq |x_0 - y_0|^2 e^{-K/2T} + (2\Gamma/K + C_{\mathcal{R}})h^{\beta}.$$

3.2.3 Optimal choice of parameters

Theorem 3.4 is fundamental in terms of applying the MLMC as it guarantees that the estimate for the variance in (7) holds. In particular, we have the following lemma.

Lemma 3.7 Assume that all the assumptions from Theorem 3.4 hold. Let $g(\cdot)$ be a Lipschitz function. Define

$$h_\ell = 2^{-\ell}, \quad T_\ell \sim -\frac{2\beta}{K} (\log h_0 + \ell \log 2), \quad \forall \ell \geq 0.$$

Then, resulting MLMC variance is given by.

$$\text{Var}[\Delta_\ell] \leq 2^{-\beta\ell}, \quad \Delta_\ell = g\left(x_{\frac{T_{\ell-1}}{h_{\ell-1}}}^{(f,\ell)}\right) - g\left(x_{\frac{T_{\ell-1}}{h_{\ell-1}}}^{(c,\ell)}\right)$$

Proof Since g is a Lipschitz function and

$$\mathbb{E} \left| x_{\frac{h_\ell}{T_\ell - T_{\ell-1}}} - x_0 \right|^2 < \infty,$$

the proof is a direct consequence of Theorem 3.4. \square

Remark 3.8 Unlike in the standard MLMC complexity theorem (Giles 2015) where the cost of simulating single path is of order $\mathcal{O}(h_\ell^{-1})$, here we have $\mathcal{O}(h_\ell^{-1} |\log h_\ell|)$. This is due to the fact that terminal times are increasing with levels. For the case $h_\ell = 2^{-\ell}$ this results in cost per path $\mathcal{O}(2^{-\ell} \ell)$ and does not exactly fit the complexity theorem in Giles (2015). Clearly in the case when MLMC variance decays with $\beta > 1$, we still recover the optimal complexity of order $\mathcal{O}(\varepsilon^{-2})$. However, in the case $\beta = 1$ following the proof by Giles (2015), one can see that the complexity becomes $\mathcal{O}(\varepsilon^{-2} |\log \varepsilon|^3)$.

Remark 3.9 In the proof above we have assumed that K is independent of h , while we have also used crude bounds in order not to deal directly with all the individual constants, since these would be dependent on the numerical schemes used.

Example 3.10 In the case of the Euler–Maruyama method as we see from the analysis⁵ in Sect. 4.1 $K = 2m' - L^2 h_\ell$, $\beta = 2$, while $\mathcal{R}_n = 0$, $\mathcal{H}_n = L$. Here L is the Lipschitz constant of the drift $\nabla U(x)$.

4 Examples of suitable methods

In this section we present two (out of many) numerical schemes that fulfil the conditions of Theorem 3.4. In particular, we need to verify that our scheme is L^2 -regular in time, it has bounded numerical moments as in **H0**, and finally that it satisfies the one-step error estimates (27)–(28). Note that for both methods discussed in this section we verify condition (25) with h^2 instead of h . However, since in (25) we consider $h \in (0, 1)$, both (35) and (42) imply (25).

4.1 Euler–Maruyama method

We start by considering the explicit Euler scheme

$$S_{h,\xi}^f(x) = x + h \nabla U(x) + \sqrt{2h} \xi, \quad (31)$$

while $S^f = S^c$, that is, we are using the same numerical method for the fine and coarse paths. In order to be able to recover the integrability and regularity conditions, we will

need to impose further assumptions on the potential⁶ U . In particular, additionally to Assumption **HU0**, we assume that

HU1 There exists constant L such that for any $x, y \in \mathbb{R}^d$

$$|\nabla U(x) - \nabla U(y)| \leq L|x - y|$$

As a consequence of this assumption, we have

$$|\nabla U(x)| \leq L|x| + |\nabla U(0)| \quad (32)$$

We can now prove the L^2 -regularity in time of the scheme.

L^2 -regularity Since regularity is a property of the numerical scheme itself and it does not relate with the coupling between fine and coarse levels, for simplicity of notation we prove things directly for

$$x_{n+1,x_n} = S_{h,\xi_{n+1}}^f(x_n). \quad (33)$$

In particular, the following lemma holds.

Lemma 4.1 (L^2 -regularity) *Let **HU0** and **HU1** hold. Then, the explicit Euler scheme is L^2 -regular, i.e.*

$$\mathbb{E}_{n-1}[|x_{n,x_{n-1}} - x_{n,y_{n-1}}|^2] \leq (1 - (2m - L^2 h)h)|x_{n-1} - y_{n-1}|^2 \quad (34)$$

$$\mathbb{E}_{n-1}[|Z_n|^2] \leq h^2 L^2 |x_{n-1} - y_{n-1}|^2 \quad (35)$$

Proof The difference between the Euler scheme taking values x_{n-1} and y_{n-1} at time $n - 1$ is given by

$$\begin{aligned} x_{n,x_{n-1}} - x_{n,y_{n-1}} &= x_{n-1} - y_{n-1} \\ &\quad + h(\nabla U(x_{n-1}) - \nabla U(y_{n-1})). \end{aligned}$$

This, along with **HU0** and **HU1**, leads to

$$\begin{aligned} \mathbb{E}_{n-1}[(x_{n,x_{n-1}} - x_{n,y_{n-1}})^2] &= |x_{n-1} - y_{n-1}|^2 \\ &\quad + 2h \langle \nabla U(x_{n-1}) - \nabla U(y_{n-1}), x_{n-1} - y_{n-1} \rangle \\ &\quad + |\nabla U(x_{n-1}) - \nabla U(y_{n-1})|^2 h^2 \\ &\leq |y_{n-1} - x_{n-1}|^2 (1 - 2mh + L^2 h^2) \\ &= |y_{n-1} - x_{n-1}|^2 (1 - (2m - L^2 h)h). \end{aligned}$$

This proves the first part of the lemma. Next, due to **HU1**

$$\begin{aligned} \mathbb{E}_{n-1}[|Z_n|^2] &= h^2 \mathbb{E}_{n-1}[|\nabla U(x_{n-1}) - \nabla U(y_{n-1})|^2] \\ &\leq h^2 L^2 |x_{n-1} - y_{n-1}|^2. \end{aligned}$$

\square

⁵ As we will see there $m' \leq m$ depending on the size of $\nabla U(0)$.

⁶ This restriction will be alleviated in Sect. 4.2 by means of more advanced integrators.

Integrability In the Lipschitz case we only require mean square integrability. This will become apparent when we analyse the one-step error and (27) and (28) will hold with $p = 1$.

Lemma 4.2 (Integrability) *Let **HU0** and **HU1** hold. Then,*

$$\mathbb{E}[|x_n|^2] \leq \mathbb{E}|x_0|^2 \exp\{-(2m' - L^2h)nh\} + 2(b|\nabla U(0)|^2 + h) \frac{1 - \exp\{-(2m' - L^2h)nh\}}{(2m' - L^2h)h}$$

Proof We have

$$\begin{aligned} |x_n|^2 &= |x_{n-1}|^2 + |\nabla U(x_{n-1})|^2 h^2 + 2h\xi^T \xi \\ &\quad + 2hx_{n-1}^T \nabla U(x_{n-1}) + \sqrt{2h}x_{n-1}^T \xi \\ &\quad + \sqrt{2h}^{3/2} \xi^T \nabla U(x_{n-1}) \end{aligned}$$

and hence, using (11)

$$\mathbb{E}|x_n|^2 \leq \mathbb{E}|x_{n-1}|^2(1 - 2m'h + L^2h^2) + 2b|\nabla U(0)|^2 + 2dh.$$

We can now use Lemma 3.5

$$\begin{aligned} \mathbb{E}|x_n|^2 &\leq \mathbb{E}|x_0|^2 \exp\{-(2m' - L^2h)nh\} \\ &\quad + 2(b|\nabla U(0)|^2 + dh) \frac{1 - \exp\{-(2m' - L^2h)nh\}}{(2m' - L^2h)h} \end{aligned}$$

The proof for $q > 2$ can be done in similar way by using the binomial theorem. \square

One-step errors estimates Having proved L^2 -regularity and integrability for the Euler scheme, we are now left with the task of proving inequalities (27) and (28) for Euler schemes coupled as in Algorithm 1. It is enough to prove the results for $n = 1$. We note that both $x_0^f = x_0^c = x$ and we have the following lemma.

Lemma 4.3 (One-step errors) *Let **HU0** and **HU1** hold. Then, the weak one-step distance between Euler schemes with time steps $\frac{h}{2}$ and h , respectively, is given by*

$$\begin{aligned} |\mathbb{E}[x_{1,x}^f - x_{1,x}^c]| &\leq \frac{h^{3/2}}{2} L \left(\mathbb{E} \left[\frac{\sqrt{h}}{2} (L|x| + |\nabla U(0)|) \right] + \sqrt{\frac{2d}{\pi}} \right). \end{aligned} \quad (36)$$

The one-step L^2 distance can be estimated as

$$\mathbb{E}|x_{1,x}^f - x_{1,x}^c|^2 \leq h^3 \frac{L^2}{4} \left(\frac{h}{2} (|x|^2 + |\nabla U(0)|^2) + d \right) \quad (37)$$

If in addition to **HU0** and **HU1**, $U \in C^3$ and⁷

$$|\partial^2 U(x)| + |\partial^3 U(x)| \leq C, \quad \forall x \in \mathbb{R}^d,$$

then the order in h of the weak error bound can be improved, i.e.

$$\begin{aligned} |\mathbb{E}[x_{1,x}^f - x_{1,x}^c]| &\leq Ch^2 \mathbb{E}[|x| + h|x|^2 \\ &\quad + |\nabla U(0)| + h|\nabla U(0)|^2 + d]. \end{aligned} \quad (38)$$

Proof We calculate

$$\begin{aligned} x_{1,x}^f - x_{1,x}^c &= x + \frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 + \frac{h}{2} \nabla U \left(x + \frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right) \\ &\quad + \sqrt{h} \xi_2 - \left(x + h \nabla U(x) + \sqrt{h} (\xi_1 + \xi_2) \right) \\ &= \frac{h}{2} \nabla U \left(x + \frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right) - \frac{h}{2} \nabla U(x). \end{aligned} \quad (39)$$

It then follows from **HU1** that

$$|\mathbb{E}[x_{1,x}^f - x_{1,x}^c]| \leq \frac{h^{3/2}}{2} L \mathbb{E} \left| \frac{\sqrt{h}}{2} \nabla U(x) + \xi_1 \right|.$$

Furthermore, if we use (32), the triangle equality and the fact that $\mathbb{E}|\xi_1| = \sqrt{\frac{2d}{\pi}}$, we obtain (36). If we now assume that $U \in C^3$, then for $\delta_t = x + t(\frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1)$, $t \in [0, 1]$, we write

$$\begin{aligned} \nabla U \left(x + \frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right) &= \nabla U(x) \\ &\quad + \sum_{|\alpha|=1} \partial^\alpha \nabla U(x) \left(\frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right)^\alpha \\ &\quad + \sum_{|\alpha|=2} \int_0^1 (1-t) \partial^\alpha \nabla U(\delta_t) dt \left(\frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right)^\alpha, \end{aligned}$$

where we used multi-index notation. Consequently,

$$\begin{aligned} \mathbb{E} \left[\nabla U \left(x + \frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right) - \nabla U(x) \right] &\leq Ch^2 \mathbb{E}[(|x| + h|x|^2 + |\nabla U(0)| + h|\nabla U(0)|^2 + |\xi_1|^2)], \end{aligned}$$

which, together with $\mathbb{E}[|\xi_1|^2] = d$, gives (38). Equation (37) trivially follows from (39) by observing that

⁷ Thanks to the integrability conditions we could easily extend the analysis to the case where the derivatives are bounded by a polynomial of x .

$$\begin{aligned} & \mathbb{E}|x_{1,x}^f - x_{1,x}^c|^2 \\ & \leq L^2 \frac{h^2}{4} \mathbb{E} \left| \frac{h}{2} \nabla U(x) + \sqrt{h} \xi_1 \right|^2 \\ & \leq h^3 \frac{L^2}{4} \left(\frac{h}{2} (|x|^2 + |\nabla U(0)|^2) + d \right) \end{aligned}$$

□

Remark 4.4 In the case of log-concave target the bias of MLMC using the Euler method can be explicitly quantified using the results from Durmus and Moulines (2016).

4.2 Non-Lipschitz setting

In the previous subsection we found out that in order to analyse the regularity and the one-step error of the explicit Euler approximation, we had to impose an additional assumption about $\nabla U(x)$ being globally Lipschitz. This is necessary since in the absence of this condition, Euler method is shown to be transient or even divergent (Roberts and Tweedie 1996; Hutzenthaler et al. 2018). However, in many applications of interest this is a rather restricting condition. An example of this is the potential⁸

$$U(x) = -\frac{x^4}{4} - \frac{x^2}{2}.$$

A standard way to deal with this is to use either an implicit scheme or specially designed explicit schemes (Hutzenthaler and Jentzen 2015; Szpruch and Zhang 2018). Here we will study only the case of implicit Euler.

4.2.1 Implicit Euler method

Here we will focus on the implicit Euler scheme

$$x_n = x_{n-1} + h \nabla U(x_n) + \sqrt{2h} \xi_n$$

We will assume that Assumption **HU0** holds and moreover replace **HU1** with

HU1' Let $k \geq 1$. For any $x, y \in \mathbb{R}^d$ there exists constant L s.t

$$|\nabla U(x) - \nabla U(y)| \leq L(1 + |x|^{k-1} + |y|^{k-1})|x - y|$$

As a consequence of this assumption, we have

$$|\nabla U(x)| \leq L|x|^k + |\nabla U(0)| \quad (40)$$

⁸ One also may consider the case of products of distribution functions, where after taking the log one ends up with a polynomial in the different variables.

Integrability Uniform-in-time bounds on the p th moments of x_n for all $p \geq 1$ can be easily deduced from the results in Mao and Szpruch (2013a, b). Nevertheless, for the convenience of the reader we will present the analysis of the regularity of the scheme, where the effect of the implicitness of the scheme on the regularity should become quickly apparent.

L^2 -regularity

Lemma 4.5 (L^2 -regularity) Let **HU0** and **HU1'** hold. Then, an implicit Euler scheme is L^2 -regular, i.e.

$$\begin{aligned} \mathbb{E}_{n-1}[(x_{n,x_{n-1}} - x_{n,y_{n-1}})^2] & \leq (1 - 2mh)(y_{n-1} - x_{n-1})^2 \\ & + \mathcal{R}_{n-1}, \end{aligned} \quad (41)$$

and

$$\sum_{k=0}^{\infty} \mathcal{R}_k \leq 0.$$

Moreover,

$$\mathbb{E}_{n-1}[|Z_n|^2] \leq h^2 \mathcal{H}_{n-1} |x_{n-1} - y_{n-1}|^2, \quad (42)$$

where \mathcal{H}_{n-1} is defined by (43).

Proof The difference between the implicit Euler scheme taking values x_{n-1} and y_{n-1} time $n - 1$ is given by

$$x_{n,x_{n-1}} - x_{n,y_{n-1}} = x_{n-1} - y_{n-1} + h(\nabla U(x_n) - \nabla U(y_n)).$$

This, along with **HU0** and **HU1**, leads to

$$\begin{aligned} |x_{n,x_{n-1}} - x_{n,y_{n-1}}|^2 & = |x_{n-1} - y_{n-1}|^2 \\ & + 2h \langle \nabla U(x_n) - \nabla U(y_n), x_n - y_n \rangle \\ & - |\nabla U(x_n) - \nabla U(y_n)|^2 h^2 \\ & \leq |x_{n-1} - y_{n-1}|^2 - 2mh |x_{n,x_{n-1}} - x_{n,y_{n-1}}|^2 \end{aligned}$$

This implies

$$\begin{aligned} |x_{n,x_{n-1}} - x_{n,y_{n-1}}|^2 & \leq |x_{n-1} - y_{n-1}|^2 \frac{1}{1 + 2mh} \\ & \leq |x_{n-1} - y_{n-1}|^2 \left(1 - \frac{2mh}{1 + 2mh} \right). \end{aligned}$$

Next, we take

$$\begin{aligned} |x_{n,x_{n-1}} - y_{n,y_{n-1}}|^2 & \leq |x_{n-1} - y_{n-1}|^2 - 2mh |x_n - y_n|^2 \\ & = (1 - 2mh) |x_{n-1} - y_{n-1}|^2 \\ & - 2mh (|x_n - y_n|^2 - |x_{n-1} - y_{n-1}|^2). \end{aligned}$$

In view of Definition 3.3, we define

$$\mathcal{R}_k := -2mh(|x_k - y_k|^2 - |x_{k-1} - y_{k-1}|^2),$$

and notice that

$$\sum_{k=1}^n \mathcal{R}_k = -2mh|x_n - y_n|^2 \leq 0.$$

Hence, the proof of the first statement in the lemma is completed. Now, due to **HU1'**

$$\begin{aligned} |Z_n|^2 &= h^2 |\nabla U(x_n) - \nabla U(y_n)|^2 \\ &\leq h^2 L^2 (1 + |x_n|^{k-1} + |y_n|^{k-1})^2 |x_n - y_n|^2 \\ &\leq h^2 \left(1 - \frac{2mh}{1 + 2mh}\right) L^2 (1 + |x_n|^{k-1} + |y_n|^{k-1})^2 |x_{n-1} - y_{n-1}|^2. \end{aligned}$$

Observe that

$$\begin{aligned} \mathbb{E}_{n-1}[|x_n|^2] &= |x_{n-1}|^2 \\ &\quad + \mathbb{E}_{n-1}[2h \langle \nabla U(x_n), x_n \rangle - |\nabla U(x_n)|^2 h^2] + h \\ &\leq |x_{n-1}|^2 - mh|x_n|^2 + h(|\nabla U(0)|^2 + 1). \end{aligned}$$

Consequently,

$$\mathbb{E}_{n-1}[|x_n|^2] = \frac{1}{1 + mh} \left(|x_{n-1}|^2 + h(|\nabla U(0)|^2 + 1) \right).$$

Similarly, it can be shown that $\mathbb{E}_{n-1}[|x_n|^k]$ can be expressed as a function of $|x_{n-1}|^k$ for $k > 2$, cf. Mao and Szpruch (2013a, b). This in turn implies that there exists a constant $C > 0$ s.t.

$$\begin{aligned} \mathcal{H}_{n-1} &= \mathbb{E}_{n-1} \left[L^2 \left(1 - \frac{2mh}{1 + 2mh} \right) (1 + |x_n|^{k-1} + |y_n|^{k-1})^2 \right] \\ &\leq C(1 + |x_{n-1}|^{2(k-1)} + |y_{n-1}|^{2(k-1)}). \end{aligned} \quad (43)$$

Due to uniform integrability of the implicit Euler scheme, (26) holds. \square

One-step errors estimates Having established integrability, estimating the one-step error follows exactly the same line of the argument as in Lemma 4.3 and therefore, we skip it.

5 MLMC for SGLD

In this section we discuss the multi-level Monte Carlo method for Euler schemes with inaccurate (randomised) drifts. Namely, we consider

$$S_{h,\xi,\tau}(x) = x + hb(x, \tau) + \sqrt{2h}\xi, \quad (44)$$

where $b : \mathbb{R}^d \times \mathbb{R}^k \rightarrow \mathbb{R}^d$ and an \mathbb{R}^k -valued random variable τ are such that

$$\mathbb{E}[b(x, \tau)] = \nabla U(x) \text{ for any } x \in \mathbb{R}^d. \quad (45)$$

Our main application to Bayesian inference will be discussed in Sect. 5.1. Let us now take a sequence $(\tau_n)_{n=1}^\infty$ of mutually independent random variables satisfying (45). We assume that $(\tau_n)_{n=1}^\infty$ are also independent of the i.i.d. random variables $(\xi_n)_{n=1}^\infty$ with $\xi_n \sim \mathcal{N}(0, I)$. By analogy to the notation we used for the Euler scheme in (33), we will denote

$$\bar{x}_{n+1}, \bar{x}_n = S_{h,\xi_{n+1},\tau_{n+1}}^f(\bar{x}_n). \quad (46)$$

In the sequel we will perform a one-step analysis of the scheme defined in (46) by considering the random variables

$$\begin{aligned} \bar{x}_{1,x}^f &= S_{\frac{h}{2},\xi_2,\tau^{f,2}}^f \circ S_{\frac{h}{2},\xi_1,\tau^{f,1}}^f(x) \\ \bar{x}_{1,x}^c &= S_{h,\frac{1}{\sqrt{2}}(\xi_1+\xi_2),\tau^c}^c(x), \end{aligned} \quad (47)$$

where $\xi_1, \xi_2 \sim \mathcal{N}(0, I)$ and $\tau^{f,1}, \tau^{f,2}$ and τ^c are \mathbb{R}^k -valued random variables satisfying (45). In particular, $\tau^{f,1}$ and $\tau^{f,2}$ are assumed to be independent, but τ^c is not necessarily independent of $\tau^{f,1}$ and $\tau^{f,2}$. We note that in (47) we have coupled the noise between the fine and the coarse paths synchronously, i.e. as in Algorithm 1. One question that naturally occurs now is how one should choose to couple the random variables τ at different levels. In particular, in order for the condition with the telescopic sum to hold, one needs to have

$$\mathcal{L}(\tau^{f,1}) = \mathcal{L}(\tau^{f,2}) = \mathcal{L}(\tau^c). \quad (48)$$

We can of course just take τ^c independent of $\tau^{f,1}$ and $\tau^{f,2}$, but other choices are also possible; see Sect. 5.1 for the discussion in the context of the SGLD applied to Bayesian inference.

1. Set $x_0^{(f,\ell)} = x_0$, then simulate according to $S_{h_\ell,\xi,\tau}(x)$ for $\frac{T_\ell - T_{\ell-1}}{h_\ell}$ steps with independent random input;
2. set $x_0^{(c,\ell)} = x_0$ and $x_0^{(f,\ell)} = x_{\frac{T_\ell - T_{\ell-1}}{h_\ell}}^{h_\ell}$, then simulate $(x^{(f,\ell)}, x^{(c,\ell)})$ jointly according to

$$\begin{aligned} (x_{k+1}^{(f,\ell)}, x_{k+1}^{(c,\ell)}) &= (S_{h_\ell,\xi_{k,2},\tau_k^{f,2}} \circ S_{h_\ell,\xi_{k,1},\tau_k^{f,1}}(x_k^{(f,\ell)}), \\ &\quad S_{h_{\ell-1},\frac{1}{\sqrt{2}}(\xi_{k,1}+\xi_{k,2}),\tau_k^c}(x_k^{(c,\ell)})). \end{aligned}$$

3. set $k_\ell := \frac{T_\ell - 1}{h_{\ell-1}}$ and

$$\Delta_\ell^{(i)} := g\left(\left(x_{k_\ell}^{(f,\ell)}\right)^{(i)}\right) - g\left(\left(x_{k_\ell}^{(c,\ell)}\right)^{(i)}\right)$$

Algorithm 2: Coupling SGLD for $t_i \uparrow \infty$.

In order to bound the global error for our algorithm, we make the following assumptions on the function b in (44).

Assumption 2 (i) There is a constant $\bar{L} > 0$ such that for any \mathbb{R}^k -valued random variable τ satisfying (45) and for any $x, y \in \mathbb{R}^d$ we have

$$\mathbb{E}[|b(x, \tau) - b(y, \tau)|] \leq \bar{L}|x - y|. \quad (49)$$

(ii) There exist constants $\alpha_c, \sigma \geq 0$ such that for any \mathbb{R}^k -valued random variable τ satisfying (45) and for any $h > 0, x \in \mathbb{R}^d$ we have

$$\mathbb{E}[|b(x, \tau) - \nabla U(x)|^2] \leq \sigma^2(1 + |x|^2)h^{\alpha_c}. \quad (50)$$

Observe that conditions (49), (50) and Assumption **HU1** imply that for all random variables τ satisfying (45) and for all $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[|b(x, \tau)|^2] \leq \bar{L}_0(1 + |x|^2) \quad (51)$$

with $\bar{L}_0 := \sigma^2 h^{\alpha_c} + 2 \max(L^2, |\nabla U(0)|^2)$, cf. Section 2.4 in Majka et al. (2018). For a discussion on how to verify condition (50) for a subsampling scheme, see Example 2.15 in Majka et al. (2018). By following the proofs of Lemmas 4.1 and 4.2, we see that the L^2 regularity and integrability conditions proved therein hold for the randomised drift scheme given by (44) as well, under Assumptions **HU0** and (49). Hence, in order to be able to apply Theorem 3.4 to bound the global error for (46), we only have to estimate the one step errors, i.e. we need to verify conditions (27) and (28) in an analogous way to Lemma 4.3 for Euler schemes.

Lemma 5.1 Under Assumptions 2 and **HU1**, there is a constant $C_1 = C_1(h, x) > 0$ given by

$$C_1 := \frac{1}{4} \bar{L} \bar{L}_0^{1/2} h^{1/2} (1 + |x|^2)^{1/2} + \frac{1}{2} L \sqrt{d}$$

such that for all $h > 0$ we have

$$\mathbb{E}[\bar{x}_{1,x}^f - \bar{x}_{1,x}^c] \leq C_1 h^{3/2}. \quad (52)$$

Moreover, under the same assumptions there is a constant $C_2 = C_2(h, x) > 0$ given by

$$C_2 := \frac{1}{4} \bar{L}^2 \bar{L}_0 h^{1+(1-\alpha_c)^+} (1 + |x|^2) + d \bar{L}^2 h^{(1-\alpha_c)^+} + 8\sigma^2(1 + |x|^2)h^{(\alpha_c-1)^+}$$

such that for all $h > 0$ we have

$$\mathbb{E}[|\bar{x}_{1,x}^f - \bar{x}_{1,x}^c|^2] \leq C_2 h^{2+\min(1, \alpha_c)}. \quad (53)$$

Proof Note that we have

$$\bar{x}_{1,x}^f - \bar{x}_{1,x}^c = x + \frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1$$

$$\begin{aligned} &+ \frac{h}{2} b\left(x + \frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1, \tau_2^f\right) \\ &+ \sqrt{h} \xi_2 - x - h b(x, \tau_1^c) - \sqrt{h} (\xi_1 + \xi_2) \\ &= \frac{h}{2} b(x, \tau_1^f) + \frac{h}{2} b\left(x + \frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1, \tau_2^f\right) \\ &- h b(x, \tau_1^c). \end{aligned} \quad (54)$$

By conditioning on all the sources of randomness except for τ_2^f and using its independence of τ_1^f and ξ_1 , we show

$$\begin{aligned} &\mathbb{E}[\bar{x}_{1,x}^f - \bar{x}_{1,x}^c] \\ &= \frac{h}{2} \mathbb{E}\left[\nabla U\left(x + \frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1\right)\right] - \frac{h}{2} \nabla U(x). \end{aligned}$$

Hence, we have

$$\mathbb{E}[\bar{x}_{1,x}^f - \bar{x}_{1,x}^c] \leq \frac{h}{2} L \mathbb{E}\left[\left|\frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1\right|\right]$$

and thus, using (51) and Jensen's inequality, we obtain (52). We now use (54) to compute

$$\begin{aligned} &\mathbb{E}[|\bar{x}_{1,x}^f - \bar{x}_{1,x}^c|^2] \\ &= h^2 \mathbb{E}\left|\frac{1}{2} b(x, \tau_1^f) + \frac{1}{2} b(x, \tau_2^f) - \frac{1}{2} b(x, \tau_2^f) \right. \\ &\quad \left. + \frac{1}{2} b\left(x + \frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1, \tau_2^f\right) - b(x, \tau_1^c)\right|^2 \\ &\leq 2h^2 \mathbb{E}\left|\frac{1}{2} b(x, \tau_1^f) + \frac{1}{2} b(x, \tau_2^f) - b(x, \tau_1^c)\right|^2 \\ &\quad + \frac{1}{2} h^2 \mathbb{E}\left|b\left(x + \frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1, \tau_2^f\right) - b(x, \tau_2^f)\right|^2 \end{aligned} \quad (55)$$

Observe now that due to condition (49) the second term above can be bounded by

$$\begin{aligned} &\frac{1}{2} h^2 \bar{L}^2 \mathbb{E}\left|\frac{h}{2} b(x, \tau_1^f) + \sqrt{h} \xi_1\right|^2 \\ &\leq \frac{1}{2} h^2 \bar{L}^2 \left(\frac{h^2}{2} \mathbb{E}|b(x, \tau_1^f)|^2 + 2h \mathbb{E}|\xi_1|^2\right) \\ &\leq \frac{1}{2} h^2 \bar{L}^2 \left(\frac{h^2}{2} \bar{L}_0(1 + |x|^2) + 2hd\right), \end{aligned}$$

where in the last inequality we used (51). Moreover, the first term on the right-hand side of (55) is equal to

$$\begin{aligned} &2h^2 \mathbb{E}\left|\frac{1}{2} b(x, \tau_1^f) - \frac{1}{2} \nabla U(x) + \frac{1}{2} \nabla U(x) - \frac{1}{2} b(x, \tau_1^c) \right. \\ &\quad \left. + \frac{1}{2} b(x, \tau_2^f) - \frac{1}{2} \nabla U(x) + \frac{1}{2} \nabla U(x) - \frac{1}{2} b(x, \tau_1^c)\right|^2 \\ &\leq 2h^2 \left(\mathbb{E}|b(x, \tau_1^f) - \nabla U(x)|^2 + 2\mathbb{E}|b(x, \tau_1^c) - \nabla U(x)|^2\right) \end{aligned}$$

$$+ \mathbb{E}|b(x, \tau_2^f) - \nabla U(x)|^2) \\ \leq 8\sigma^2(1 + |x|^2)h^{2+\alpha_c},$$

where in the last inequality we used (50). This finishes the proof of (53). \square

Corollary 5.2 *If $\alpha_c = 0$ in (50), then Algorithm 2 based on the coupling given in Eq. (47) with appropriately chosen t_i has complexity $\varepsilon^{-2}|\log(\varepsilon)|^3$. If $\alpha_c > 0$, then the algorithm has complexity ε^{-2} .*

Proof Because of Lemma 5.1 we can apply the results of Sect. 3.2. In particular, if we choose T_ℓ according to Lemma 3.7 we thus for $\alpha_c = 0$ have $\beta = 1$ in Theorem 3.4 and then the complexity follows from Remark 3.8. Similarly, for $\alpha_c > 0$ we have $\beta > 1$ and Remark 3.8 concludes the proof. \square

5.1 Bayesian inference using MLMC for SGLD

The main computational task in Bayesian statistics is the approximation of expectations with respect to the posterior. The a priori uncertainty in a parameter x is modelled using a probability density $\pi_0(x)$ called the prior. Here, we consider the case where for a fixed parameter x the data $\{y_i\}_{i=1,\dots,N}$ are supposed to be i.i.d. with density $\pi(y|x)$. By Bayes' rule the posterior is given by

$$\pi(x) \propto \pi_0(x) \prod_{i=1}^N \pi(y_i|x).$$

This distribution is invariant for the Langevin equation (2) with

$$\nabla U(x) = \nabla \log \pi_0(x) + \sum_{i=1}^N \nabla \log \pi(y_i|x). \quad (56)$$

Provided that appropriate assumptions are satisfied for U , we can thus use Algorithm 1 with Euler or implicit Euler schemes to approximate expectations with respect to $\pi(dx)$. For large N the sum in Eq. (56) becomes a computational bottleneck. One way to deal with this is to replace the gradient by a lower cost stochastic approximation. In the following we apply our MLMC for SGLD framework to the recursion in Eq. (4)

$$x_{k+1} = x_k + h \left(\nabla \log \pi_0(x_k) + \frac{N}{s} \sum_{i=1}^s \nabla \log \pi(y_{\tau_i^k}|x_k) \right) \\ + \sqrt{2h}\xi_k,$$

where we take $\tau_i^k \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(\{1, \dots, N\})$ for $i = 1, \dots, s$ where by $\mathcal{U}(\{1, \dots, N\})$ we denote the uniform distribution

on $1, \dots, N$ which corresponds to sampling s items with replacement from $1, \dots, N$. Notice that each step only costs s instead of N . We make the following assumptions on the densities $\pi(y|x)$ and $\pi_0(x)$.

Assumption 3 (i) Lipschitz conditions for prior and likelihood: There exist constants $L_0, L_1 > 0$ such that for all i, x, y

$$|\nabla \log \pi(y_i|x) - \nabla \log \pi(y_i|y)| \leq L_1|x - y| \\ |\nabla \log \pi_0(x) - \nabla \log \pi_0(y)| \leq L_0|x - y|.$$

(ii) Convexity conditions for prior and likelihood: There exist constants $m_0 \geq 0$ and $m_{y_i} \geq 0$ for $i = 1, \dots, N$ such that for all i, x, y

$$\log \pi_0(y) \leq \log \pi_0(x) + \langle \nabla \log \pi_0(x), y - x \rangle \\ - \frac{m_0}{2}|x - y|^2 \\ \log \pi(y_i|y) \leq \log \pi(y_i|x) \\ + \langle \nabla \log \pi(y_i|x), y - x \rangle - \frac{m_{y_i}}{2}|x - y|^2$$

$$\text{with } \inf_i(m_0 + m_{y_i}) > 0.$$

We note that these conditions imply that the scheme given by (47) with

$$b(x, \tau) := \nabla \log \pi_0(x) + \frac{N}{s} \sum_{i=1}^s \nabla \log \pi(y_{\tau_i}|x)$$

for $x \in \mathbb{R}^d$, $\tau \in \mathbb{R}^s$, satisfies Assumptions **HU0**, **HU1** and (49). The value of the variance σ of the estimator of the drift in (50) depends on the number of samples s , cf. Example 2.15 in Majka et al. (2018).

Regarding the coupling of $\tau^{f,1}$, $\tau^{f,2}$ and τ^c , we have several possible choices. We first take s independent samples $\tau^{f,1}$ on the first fine step and another s independent samples $\tau^{f,2}$ on the second fine step. The following three choices of τ^c ensure that Eq. (48) holds.

- (i) an independent sample of $\{1, \dots, N\}$ without replacement denoted as τ_{ind}^c and called independent coupling;
- (ii) a draw of s samples without replacement from $(\tau^{f,1}, \tau^{f,2})$ denoted as τ_{union}^c and called union coupling;
- (iii) the concatenation of a draw of $\frac{s}{2}$ samples without replacement from $\tau^{f,1}$ and a draw of $\frac{s}{2}$ samples without replacement from $\tau^{f,2}$ (provided that s is even) denoted as τ_{strat}^c and called stratified coupling.

We stress that any of these couplings can be used in Algorithm 2. The problem of coupling the random variables τ between different levels in an optimal way will be further investigated in our future work.

6 Numerical investigations

In this section we perform numerical simulations that illustrate our theoretical findings. We start by studying an Ornstein–Uhlenbeck process in Sect. 6.1 using the explicit Euler method, while in Sect. 6.3 we study a Bayesian logistic regression model using the SGLD.

6.1 Ornstein Uhlenbeck process

We consider the SDE

$$dX_t = -\kappa X_t dt + \sqrt{2}dW_t, \quad (57)$$

and its discretisation using the Euler method

$$x_{n+1} = S_{h,\xi}(x_n), \quad S_{h,\xi}(x) = x - h\kappa x + \sqrt{2h}\xi. \quad (58)$$

Equation (57) is ergodic with its invariant measure being $\mathcal{N}(0, \kappa^{-1})$. Furthermore, it is possible to show that the Euler method (58) is similarly ergodic with its invariant measure (Zygalakis 2011) being $\mathcal{N}\left(0, \frac{2}{2\kappa - \kappa^2 h}\right)$. In Fig. 2, we plot the outputs of our numerical simulations using Algorithm 1. The parameter of interest here is the variance of the invariant measure κ^{-1} which we try to approximate for different mean square error tolerances ε .

More precisely, in Fig. 2a we see the allocation of samples for various levels with respect to ε , while in Fig. 2b we compare the computational cost of the algorithm as a function of the parameter ε . As we can see, the computational complexity grows as $\mathcal{O}(\varepsilon^{-2})$ as predicted by our theory [here $\alpha = \beta = 2$ in (27) and (28)].

Finally, in Fig. 2c we plot the approximation of the variance κ^{-1} from our algorithm. Note that this coincides with the choice $g(x) = x^2$ since the mean of the invariant measure is 0. As we can see, as ε becomes smaller, even though the estimator is in principle biased, we get perfect agreement with the true value of the variance.

6.2 Non-Lipschitz

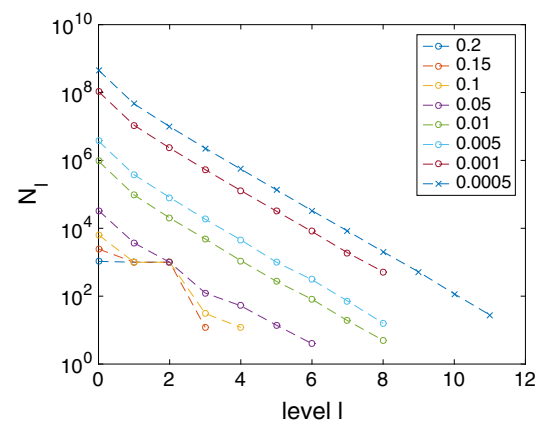
We consider the SDE

$$dX_t = -\left(X_t^3 + X_t\right)dt + \sqrt{2}dW_t, \quad (59)$$

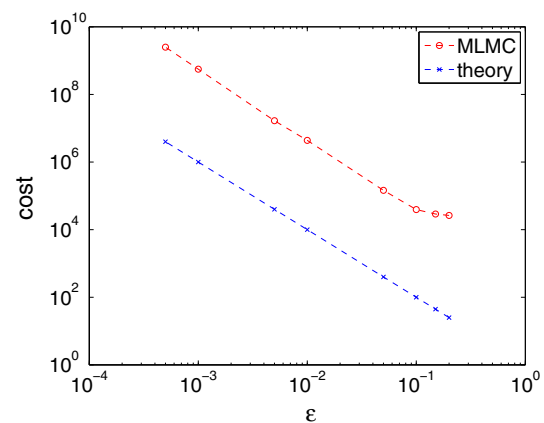
and its discretisation using the implicit Euler method

$$x_{n+1} = x_n - h\left(x_{n+1}^3 + x_{n+1}\right) + \sqrt{2h}\xi_n. \quad (60)$$

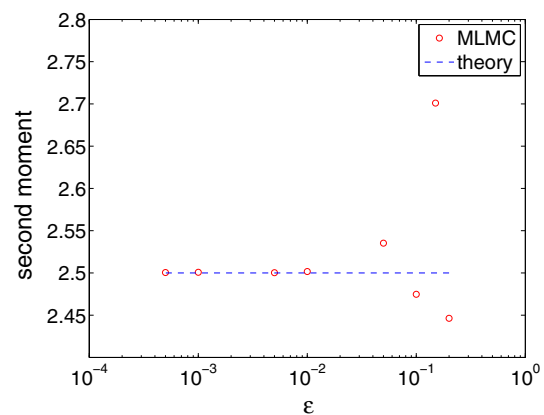
In Fig. 3, we plot the outputs of our numerical simulations using Algorithm 1. The parameter of interest here is the second moment of the invariant measure $\int_{\mathbb{R}} x^2 \exp$



(a) Numbers of samples on different levels for given accuracy



(b) Cost vs accuracy



(c) Estimation of κ^{-1}

Fig. 2 MLMC results for (57) for $g(x) = x^2$ and $\kappa = 0.4$

$\left(-\frac{1}{4}x^4 - \frac{1}{2}x^2\right)dx$ which we try to approximate for different mean square error tolerances ε .

More precisely, in Fig. 3a we see the allocation of samples for various levels with respect to ε , while in Fig. 3b we compare the computational cost of the algorithm as a function

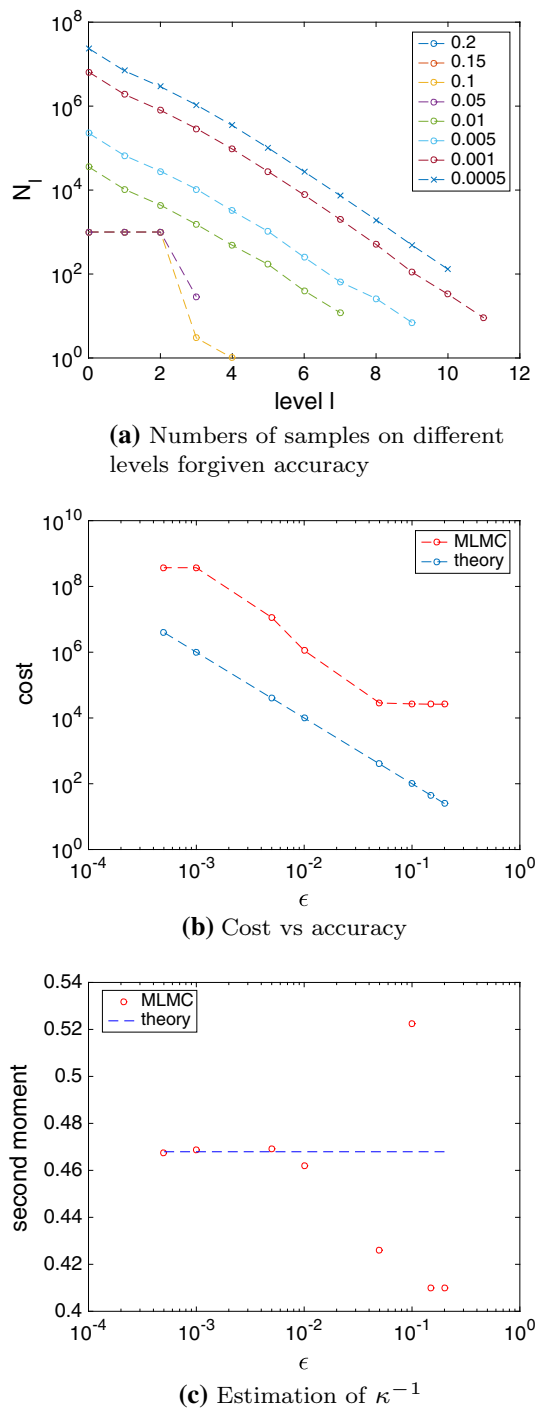


Fig. 3 MLMC results for (59) for $g(x) = x^2$

of the parameter ϵ . As we can see, the computational complexity grows as $\mathcal{O}(\epsilon^{-2})$ as predicted by our theory [here $\alpha = \beta = 2$ in (27) and (28)].

Finally, in Fig. 3c we plot the approximation of the second moment of the invariant measure from our algorithm. As we can see as ϵ becomes smaller, even though the estimator is

in principle biased, we get perfect agreement with the true value⁹ of the second moment.

6.3 Bayesian logistic regression

In the following we present numerical simulations for a binary Bayesian logistic regression model. In this case the data $y_i \in \{-1, 1\}$ are modelled by

$$p(y_i | \iota_i, x) = f(y_i x^T \iota_i) \quad (61)$$

where $f(z) = \frac{1}{1 + \exp(-z)} \in [0, 1]$ and $\iota_i \in \mathbb{R}^d$ are fixed covariates. We put a Gaussian prior $\mathcal{N}(0, C_0)$ on x , for simplicity we use $C_0 = I$ subsequently. By Bayes' rule the posterior density $\pi(x)$ satisfies

$$\pi(x) \propto \exp\left(-\frac{1}{2}|x|_{C_0}^2\right) \prod_{i=1}^N f(y_i x^T \iota_i).$$

We consider $d = 3$ and $N = 100$ data points and choose the covariate to be

$$\iota = \begin{pmatrix} \iota_{1,1} & \iota_{1,2} & 1 \\ \iota_{2,1} & \iota_{2,2} & 1 \\ \vdots & \vdots & \vdots \\ \iota_{100,1} & \iota_{100,2} & 1 \end{pmatrix}$$

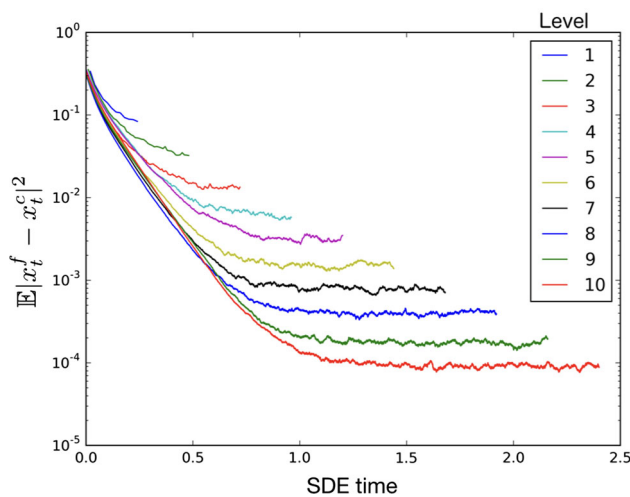
for a fixed sample of $\iota_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \dots, 100$ and $j = 1, 2$.

In Algorithm 2 we can choose the starting position x_0 . It is reasonable to start the path of the individual SGLD trajectories at the mode of the target distribution (heuristically this makes the distance $\mathbb{E}|x_0^{(c,\ell)} - x_0^{(f,\ell)}|$ in step 2 in Algorithm 2 small). That is, we set the x_0 to be the maximum a posteriori estimator (MAP)

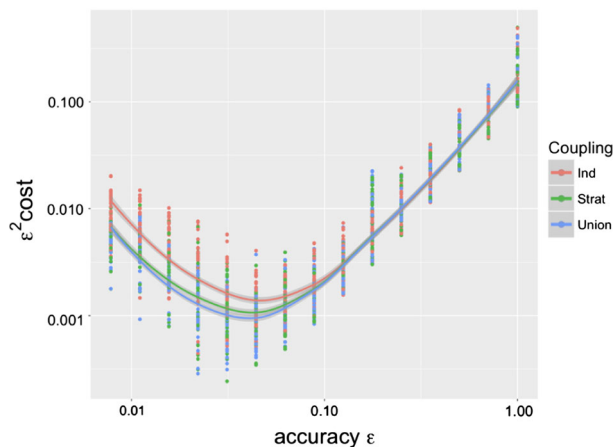
$$x_0 = \operatorname{argmax}_x \exp\left(-\frac{1}{2}|x|_{C_0}^2\right) \prod_{i=1}^N f(y_i x^T \iota_i)$$

which is approximated using the Newton–Raphson method. Our numerical results are described in Fig. 4. In particular, in Fig. 4a we illustrate the behaviour of the coupling by plotting an estimate of the average distance during the joint evolution in step 2 of Algorithm 2. The behaviour in this figure agrees qualitatively with the statement of Theorem 3.4, as T grows, there is an initial exponential decay up to an additive constant. For the simulation we used $h_0 = 0.02$, $T_\ell = 3(\ell + 1)$ and $s = 20$. Furthermore, in Fig. 4b we plot CPU-time $\times \epsilon^2$ against ϵ for the estimation of the mean. The objective here is to estimate the mean square distance from the MAP estimator

⁹ Which has been calculated using high-order quadrature.



(a) Coupling difference.



(b) Cost of MLMC.

Fig. 4 **a** Illustration of the joint evolution in step 2 of Algorithm 2 for the union coupling, **b** cost of MLMC (sequential CPU time) SGLD for Bayesian logistic regression for decreasing accuracy parameter ε and different couplings

x_0 and the posterior that is $\int |x - x_0|^2 \pi(x) dx$. Again, after some initial transient where CPU-time $\times \varepsilon^2$ decreases, we see that we get a quantitative agreement with our theory since the CPU-time $\times \varepsilon^2$ increases in a logarithmic way in the limit of ε going to zero.

Acknowledgements MBM is supported by the EPSRC Grant EP/P003818/1. KCZ is partially supported by the Alan Turing Institute under the EPSRC Grant EP/N510129/1. We would like to thank the referees for their careful reading of our manuscript and for numerous useful suggestions.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Abdulle, A., Vilmart, G., Zygalkis, K.C.: High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM J. Numer. Anal.* **52**(4), 1600–1622 (2014). <https://doi.org/10.1137/130935616>
- Agapiou, S., Roberts, G.O., Vollmer, S.J.: Unbiased Monte Carlo: posterior estimation for intractable/infinite-dimensional models. *Bernoulli* **24**(3), 1726–1786 (2018). <https://doi.org/10.3150/16-BEJ911>
- Duffie, D., Glynn, P.: Efficient Monte Carlo simulation of security prices. *Ann. Appl. Probab.* **5**(4), 897–905 (1995)
- Durmus, A., Moulines, E.: High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm (2016). arXiv e-prints [arXiv:1605.01559](https://arxiv.org/abs/1605.01559)
- Fang, W., Giles, M.B.: Adaptive Euler–Maruyama method for SDEs with non-globally Lipschitz drift: Part I, finite time interval (2016). arXiv preprint [arXiv:1609.08101](https://arxiv.org/abs/1609.08101)
- Fang, W., Giles, M.B.: Adaptive Euler–Maruyama method for SDEs with non-globally Lipschitz drift: Part II, infinite time interval (2017). arXiv preprint [arXiv:1703.06743](https://arxiv.org/abs/1703.06743)
- Fang, W., Giles, M.B.: Multilevel Monte Carlo method for ergodic SDEs without contractivity. *J. Math. Anal. Appl.* **476**(1), 149–176 (2019). <https://doi.org/10.1016/j.jmaa.2018.12.032>
- Giles, M.B.: Multilevel Monte Carlo methods. *Acta Numer.* **24**, 259–328 (2015). <https://doi.org/10.1017/S096249291500001X>
- Giles, M.B., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *Ann. Appl. Probab.* **24**(4), 1585–1620 (2014). <https://doi.org/10.1214/13-AAP957>
- Glynn, P.W., Rhee, C.H.: Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* **51**, 377–389 (2014). <https://doi.org/10.1239/jap/1417528487> (Celebrating 50 Years of The Applied Probability Trust)
- Hairer, M., Mattingly, J.C., Scheutzow, M.: Asymptotic coupling and a general form of Harris’ theorem with applications to stochastic delay equations. *Probab. Theory Relat. Fields* **149**(1–2), 223–259 (2011). <https://doi.org/10.1007/s00440-009-0250-6>
- Has’minskiĭ, R.Z.: Stochastic Stability of Differential Equations, Monographs and Textbooks on Mechanics of Solids and Fluids: Mechanics and Analysis, vol/ 7. Sijthoff & Noordhoff, Alphen aan den Rijn—Germantown, Md., translated from the Russian by D. Louvish (1980)
- Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109 (1970). <https://doi.org/10.1093/biomet/57.1.97>
- Hutzenthaler, M., Jentzen, A.: Numerical approximations of stochastic differential equations with non-globally Lipschitz continuous coefficients. *Mem. Am. Math. Soc.* **236**(1112), v+99 (2015). <https://doi.org/10.1090/memo/1112>
- Hutzenthaler, M., Jentzen, A., Kloeden, P.E.: Strong and weak divergence in finite time of Euler’s method for stochastic differential equations with non-globally Lipschitz continuous coefficients. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **467**(2130), 1563–1576 (2011). <https://doi.org/10.1098/rspa.2010.0348>
- Hutzenthaler, M., Jentzen, A., Wang, X.: Exponential integrability properties of numerical approximation processes for nonlinear stochastic differential equations. *Math. Comput.* **87**(311), 1353–1413 (2018). <https://doi.org/10.1090/mcom/3146>
- Kebaier, A.: Statistical Romberg extrapolation: a new variance reduction method and applications to option pricing. *Ann. Appl. Probab.* **15**(4), 2681–2705 (2005). <https://doi.org/10.1214/105051605000000511>
- Kloeden, P., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1992)

- Krylov, N.V.: Controlled Diffusion Processes, Stochastic Modelling and Applied Probability, vol. 14. Springer, Berlin (2009) (translated from the 1977 Russian original by A. B. Aries, Reprint of the 1980 edition)
- Lamberton, D., Pagès, G.: Recursive computation of the invariant distribution of a diffusion. *Bernoulli* **8**(3), 367–405 (2002). <https://doi.org/10.1142/S0219493703000838>
- Majka, M.B., Mijatović, A., Szpruch, L.: Non-asymptotic bounds for sampling algorithms without log-concavity (2018). arXiv e-prints [arXiv:1808.07105](https://arxiv.org/abs/1808.07105)
- Mao, X., Szpruch, L.: Strong convergence and stability of implicit numerical methods for stochastic differential equations with non-globally Lipschitz continuous coefficients. *J. Comput. Appl. Math.* **238**, 14–28 (2013a). <https://doi.org/10.1016/j.cam.2012.08.015>
- Mao, X., Szpruch, L.: Strong convergence rates for backward Euler-Maruyama method for non-linear dissipative-type stochastic differential equations with super-linear diffusion coefficients. *Stochastics* **85**(1), 144–171 (2013b). <https://doi.org/10.1080/17442508.2011.651213>
- Mattingly, J.C., Stuart, A.M., Higham, D.J.: Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stoch. Process Appl.* **101**(2), 185–232 (2002). [https://doi.org/10.1016/S0304-4149\(02\)00150-3](https://doi.org/10.1016/S0304-4149(02)00150-3)
- Nagapetyan, T., Duncan, A.B., Hasenclever, L., Vollmer, S.J., Szpruch, L., Zygalakis, K.: The True Cost of Stochastic Gradient Langevin Dynamics (2017). arXiv e-prints [arXiv:1706.02692](https://arxiv.org/abs/1706.02692)
- Rhee, C.H., Glynn, P.W.: A new approach to unbiased estimation for SDE's. In: Proceedings of the Winter Simulation Conference, Winter Simulation Conference, WSC'12, pp. 17:1–17:7 (2012). <http://dl.acm.org/citation.cfm?id=2429759.2429780>
- Rhee, C.H., Glynn, P.W.: Unbiased estimation with square root convergence for SDE models. *Oper. Res.* **63**(5), 1026–1043 (2015). <https://doi.org/10.1287/opre.2015.1404>
- Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**(4), 341–363 (1996). <https://doi.org/10.2307/3318418>
- Shardlow, T., Stuart, A.M.: A perturbation theory for ergodic Markov chains and application to numerical approximations. *SIAM J. Numer. Anal.* **37**(4), 1120–1137 (2000). <https://doi.org/10.1137/S0036142998337235>
- Szpruch, L., Zhang, X.: V-integrability, asymptotic stability and comparison property of explicit numerical schemes for non-linear SDEs. *Math. Comput.* **87**(310), 755–783 (2018). <https://doi.org/10.1090/mcom/3219>
- Talay, D.: Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process Relat. Fields* **8**(2), 163–198 (2002)
- Talay, D., Tubaro, L.: Expansion of the global error for numerical schemes solving stochastic differential equations. *Stoch. Anal. Appl.* **8**(4), 483–509 (1990). <https://doi.org/10.1080/07362999008809220>
- Teh, Y.W., Thiery, A.H., Vollmer, S.J.: Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.* **17**, Paper No. 7, 33 (2016)
- Vollmer, S.J., Zygalakis, K.C., Teh, Y.W.: Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.* **17**, Paper No. 159, 45 (2016)
- Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th ICML (2011)
- Zygalakis, K.C.: On the existence and the applications of modified equations for stochastic differential equations. *SIAM J. Sci. Comput.* **33**(1), 102–130 (2011). <https://doi.org/10.1137/090762336>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.